Ann. Inst. Statist. Math.
30 (1978), Part B, 185-197

ANALYSIS OF CROSS CLASSIFIED DATA BY AIC

YOSIYUKI SAKAMOTO AND HIROTUGU AKAIKE

(Received Mar. 19, 1977; revised Dec. 13, 1977)

Abstract

The purpose of the present paper is to propose a simple but practically useful procedure for the analysis of multidimensional contingency tables of survey data. By the procedure we can determine the predictor on which a specific variable has the strongest dependence and also the optimal combination of predictors. The procedure is very simply realized by the search for the minimum of the statistic AIC within a set of models proposed in this paper. The practical utility of the procedure is demonstrated by the results of some successful applications to the analysis of the survey data of the Japanese national character. The difference between the present procedure and the conventional test procedure is briefly discussed.

1. Introduction

Tables 1.1, 1.2 and 1.3 are a part of the survey results obtained by the 1973 nation wide survey of the Japanese national character [6], [7]. The question asked was: "On the whole in Japan, which sex do you think has the more difficult life, men or women?" These tables evidently show that the answer to this question depends most significantly on sex, among the three demographic factors.

How can we form such a judgement? How can we evaluate the strength of the dependence of the answer on those three factors?

Table 1.1

		Sex		Total
		Male (S_1)	Female (S_2)	
Which sex has more difficult life?	Men (W_1)	904	790	1694
	Women (W_2)	491	870	1361
Total		1395	1660	3055

Table 1.2

	Age				Total
	20-29 (A_1)	30-39 (A_2)	40-49 (A_3)	50 yrs & over (A_4)	
Men	454	408	363	469	1694
Women	324	319	303	415	1361
Total	778	727	666	884	3055

Table 1.3

	Rural vs. urban breakdown				Total
	6 Metropolitan cities (R_1)	Other cities		Rural (R_4)	
		Pop.: 200,000 & over (R_2)	Pop.: Under 200,000 (R_3)		
Men	322	388	525	459	1694
Women	217	304	445	395	1361
Total	539	692	970	854	3055

Perhaps we are tacitly assuming that the dependences can be evaluated by some standard which is common to all these situations.

In this paper we first propose a search procedure for the predictor on which a specific variable has the strongest dependence and then propose a procedure to search for the optimal combination of predictors. Here 'optimal' means that the combination demonstrates the most significant dependence between the variable predicted and the predictors. To solve these problems we propose the use of some models which describe the dependence relations among the variables. The discrepancy of a model fitted to a set of observed data by the method of maximum likelihood is evaluated by the statistic AIC defined by the following [1], [2]:

$$\text{AIC}=(-2)\log(\text{maximized likelihood})+2k\,, \tag{1.1}$$

where log denotes the natural logarithm and k is the number of parameters within the model which are adjusted to attain the maximum of the likelihood.

The introduction of AIC is based on the entropy maximization principle: formulate the object of statistical inference as the estimation of the true distribution from the data and try to find the estimate which will maximize the expected entropy. The entropy is a natural measure of discrimination between the true and the estimated probability distribution, $f(x)$ and $g(x:\theta)$, and is defined by

$$B(f:g(\cdot;\theta))=-\int f(x)\log\{f(x)/g(x;\theta)\}dx \tag{1.2}$$

$$= \mathrm{E} \log g(x; \theta) - \mathrm{E} \log f(x) \ .$$

A large value of the entropy $B(f : g(\cdot; \theta))$ means that the distribution $g(x; \theta)$ is a good approximation to the true distribution $f(x)$.

Consider the situation where the family of models $\bigcup_{k=1}^{L} \{g(x; {}_k\theta)\}$ is given, where $g(x; {}_k\theta)$ is specified by the vector of parameters ${}_k\theta = (\theta_1, \theta_2, \cdots, \theta_k, 0_{k+1}, \cdots, 0_L)$ $(k=1, \cdots, L)$ and it is assumed that $f(x) = g(x; {}_p\theta) = g(x; \theta_0)$ for some p $(1 \leqq p \leqq L)$. Here 0_k denotes a prescribed value of θ_k. Denote by ${}_k\hat{\theta}$ the maximum likelihood estimate of the parameter ${}_k\theta$, then the familiar log likelihood ratio statistic is given by ${}_k\eta_L = (-2)\{\log g(x; {}_k\hat{\theta}) - \log g(x; {}_L\hat{\theta})\}$ and the statistic $({}_k\eta_L + 2k - L)/n$ is an asymptotically unbiased estimate of $-\mathrm{E}\, B(g(\cdot; \theta_0), g(\cdot; {}_k\hat{\theta}))$ [1]. For the purpose of comparison of $g(x; {}_k\hat{\theta})$, the common constant $2 \log g(x; {}_L\hat{\theta}) - L$ is ignored and we get an information criterion (AIC) of (1.1). We regard a model with a smaller AIC as a better one, as it is expected to have a larger entropy. The model with the minimum AIC will be called the minimum AIC estimate or MAICE. Detailed discussions of these concepts are found in [1], [2].

To illustrate the use of AIC we consider the classical test of independence. From the point of view of the statistic AIC the conventional test of independence of a two-way contingency table $\{n(i, j): i=1, \cdots, r,\ j=1, \cdots, c\}$ is regarded as a comparison of the unrestricted model and the independence model defined by $p(i, j) = p(i, \cdot)p(\cdot, j)$, where $p(i, j)$ denotes the probability of observing a combination (i, j), $p(i, \cdot) = \sum_j p(i, j)$ and $p(\cdot, j) = \sum_i p(i, j)$. The corresponding maximum likelihood estimates of $p(i, j)$ are given by $n(i, j)/n$ and $\{n(i, \cdot)n(\cdot, j)\}/n^2$, respectively. Due to the constraint $\sum_i \sum_j p(i, j) = 1$, the number of free parameters in the first model is $(rc-1)$ and, due to the constraints $\sum_j p(\cdot, j) = 1$ and $\sum_i p(i, \cdot) = 1$, that in the second is $(r-1)+(c-1)$. The AIC's for these models are respectively given by

$$\text{(1.3)} \qquad \mathrm{AIC}_1 = (-2) \sum_i \sum_j n(i, j) \log \{n(i, j)/n\} + 2(rc-1)$$

$$\text{(1.4)} \qquad \mathrm{AIC}_0 = (-2) \sum_i \sum_j n(i, j) \log \{n(i, \cdot)n(\cdot, j)/n^2\} + 2(r+c-2) \ .$$

The definition of AIC suggests that the independence model should be adopted if AIC_0 is smaller than AIC_1, otherwise the dependence model should be adopted. If we follow this suggestion we take the MAICE as our choice. This defines the MAICE procedure.

In the case of the analysis of the Tables 1.1, 1.2 and 1.3, it would be reasonable to assume that in evaluating the dependence between the

opinion and a factor we are neglecting the effects of the remaining two factors. This idea leads us to a set of models to be defined in the next section.

2. The simplest models and their AIC's

Assume that a k-way contingency table consists of a variable to be predicted (denoted by i_1) and the $k-1$ predictors (denoted by $i_2, i_3, \cdots, i_k$). We denote the joint probability by $p(i_1, i_2, \cdots, i_k)$ and the cell frequency by $n(i_1, i_2, \cdots, i_k)$ ($\sum_{i_1, \cdots, i_k} n(i_1, i_2, \cdots, i_k)=n$), where i_j is used to represent one of the values $1, 2, \cdots, C_{i_j}$ which are taken by the variable i_j ($j=1, 2, \cdots, k$). In these representations we will simply discard a variable when a sum is taken with respct to its values. For example, we put

$$p(i_1, i_2, \cdots, i_{k-1})=\sum_{i_k} p(i_1, i_2, \cdots, i_k)$$

(2.1) and

$$n(i_1, i_2, \cdots, i_{k-2})=\sum_{i_{k-1}} n(i_1, i_2, \cdots, i_{k-1}) .$$

First consider the search for a single predictor on which the variable to be predicted has the strongest dependence. The simplest model which is in accordance with the observation at the end of the preceding section can be obtained by assuming the simplest possible structure which completely ignores the dependence between the variables left out of our consideration. This is given by

$$p(i_1, \cdots, i_k)=p(i_1, i_l) \prod_{j=2, j\neq l}^{k} p(i_j) \qquad l=2, 3, \cdots, k . \tag{2.2}$$

The log likelihood of a model belonging to (2.2) for a sample with cell frequencies $n(i_1, \cdots, i_k)$ is given by

$$L=\sum_{i_1, \cdots, i_k} n(i_1, \cdots, i_k) \log \left\{p(i_1, i_l) \prod_{j=2, j\neq l}^{k} p(i_j)\right\} .$$

By maximizing L with respect to $p(i_1, i_l)$'s and $p(i_j)$'s the maximum likelihood estimate of the joint probability is obtained by $\{n(i_1, i_l)/n^{k-1}\} \cdot \prod_{j=2, j\neq l}^{k} n(i_j)$. Since there are constraints

$$\sum_{i_1, i_l} p(i_1, i_l)=1 \qquad \text{and} \qquad \sum_{i_j} p(i_j)=1 ,$$

the model has $\{(C_{i_1}C_{i_l}-1)+\sum_{j=2, j\neq l}^{k} (C_{i_j}-1)\}$ parameters to be specified. Thus the statistic AIC for the model is given by

$$(2.3)\quad \mathrm{AIC}=(-2)\sum_{i_1,\cdots,i_k} n(i_1,\cdots,i_k)\log\left\{n(i_1,i_l)\prod_{j=2,j\neq l}^{k} n(i_j)/n^{k-1}\right\}$$
$$+2\left\{(C_{i_1}C_{i_l}-1)+\sum_{j=2,j\neq l}^{k}(C_{i_j}-1)\right\}$$
$$=(-2)\left[\sum_{i_1,i_l} n(i_1,i_l)\log\{n(i_1,i_l)/n\}+\sum_{j=2}^{k}\sum_{i_j} n(i_j)\log\{n(i_j)/n\}\right.$$
$$\left.-\sum_{i_l} n(i_l)\log\{n(i_l)/n\}\right]+2\left\{(C_{i_1}C_{i_l}-1)+\sum_{j=2}^{k}(C_{i_j}-1)-(C_{i_l}-1)\right\}.$$

For the purpose of comparison of models within the above set the common constant $(-2)\sum_{j=2}^{k}\sum_{i_j} n(i_j)\log\{n(i_j)/n\}+2\sum_{j=2}^{k}(C_{i_j}-1)$ is ignored and the statistic AIC is given by

$$(2.4)\quad \mathrm{AIC}=(-2)\left[\sum_{i_1,i_l} n(i_1,i_l)\log\{n(i_1,i_l)/n\}-\sum_{i_l} n(i_l)\log\{n(i_l)/n\}\right]$$
$$+2\{(C_{i_1}C_{i_l}-1)-(C_{i_l}-1)\}.$$

Tables 1.1, 1.2 and 1.3 of Section 1 clearly show that the response to the question 'which sex has more difficult life' depends most significantly on sex. It will be interesting to see if the MAICE procedure with the present model confirms this observation. The number of variables in our example is 4 and the necessary AIC's are obtained by putting $k=4$ in (2.3) or (2.4), where i_1 denotes a category of the opinion, i_2

Table 2

No.	Model No.	Model	AIC	No. of Parameters	χ^2-value	$1-F(\chi^2)$	Degrees of Freedom
1	(0, 1)	$p(i_1,i_2,i_3,i_4)$	25132.68	63	—	—	—
2	(1, 1)	$p(i_1,i_2,i_3)p(i_2,i_3,i_4)/p(i_2,i_3)$	25100.08	39	15.329	0.911	24
3	(1, 2)	$p(i_1,i_2,i_4)p(i_2,i_3,i_4)/p(i_2,i_4)$	25102.18	39	17.461	0.828	24
4	(1, 3)	$p(i_1,i_3,i_4)p(i_2,i_3,i_4)/p(i_3,i_4)$	25203.70	47	101.937	0.000	16
5	(2, 1)	$p(i_1,i_2)p(i_2,i_3,i_4)/p(i_2)$	25097.44*	33	24.686	0.740	30
6	(2, 2)	$p(i_1,i_3)p(i_2,i_3,i_4)/p(i_3)$	25187.96	35	110.096	0.000	28
7	(2, 3)	$p(i_1,i_4)p(i_2,i_3,i_4)/p(i_4)$	25187.20	35	109.520	0.000	28
8	(3, 1)	$p(i_1)p(i_2,i_3,i_4)$	25187.05	32	115.272	0.000	31
Model 1°		$p(i_1,i_2)p(i_3)p(i_4)$	25102.34	9	78.504	0.016	54
Model 2°		$p(i_1,i_3)p(i_2)p(i_4)$	25192.86	11	166.561	0.000	52
Model 3°		$p(i_1,i_4)p(i_2)p(i_3)$	25192.10	11	166.463	0.000	52

i_1: The question 'which sex has more difficult life?'
i_2: Sex
i_3: Age
i_4: Rural vs. urban breakdown
*: MAICE among all the models

of sex, i_3 of age and i_4 of urban vs. rural breakdown. To search for the predictor on which the opinion has the strongest dependence we have only to calculate AIC's for the two-way contingency tables shown in Tables 1.1, 1.2 and 1.3 and pick the one with the minimum AIC. From (2.3) we get 25102.34, 25192.86 and 25192.10, which are shown at the bottom of Table 2, as the AIC's of the models. Apparently the MAICE procedure suggests that we should adopt sex as the most effective predictor, which is identical to our empirical judgement. The detail of the dependence depends on how we categorize each predictor. This aspect will be discussed elsewhere [8].

3. More general models and their AIC's

A useful model for the search of an optimal combination of predictors can be obtained by using the multiplicative models of contingency tables which have previously been discussed by many authors such as Darroch [4], Bishop [3], Goodman [5], and Wermuth [9], [10]. A multiplicative model is a model such that the joint distribution of several variables is factored into the product of marginal distributions of subgroups of variables. One example of multiplicative model with $k=5$ is given by

$$(3.1)\quad p(i_1,\cdots,i_5)=p(i_1,i_4,i_5)p(i_2,i_4,i_5)p(i_3,i_4,i_5)/\{p(i_4,i_5)p(i_4,i_5)\}\ .$$

This model can be written as

$$(3.2)\qquad p(i_1,\cdots,i_5)=p(i_1|i_4,i_5)p(i_2|i_4,i_5)p(i_3|i_4,i_5)p(i_4,i_5)\ ,$$

where $p(i_1|i_4,i_5)$ denotes the conditional probability of i_1 given (i_4,i_5). This shows that in the model each of the variable pairs (i_1,i_2), (i_1,i_3) and (i_2,i_3) has zero partial association, that is, the variables in a pair is conditionally mutually independent, given the remaining three variables. Each multiplicative model is characterized by the variable groups in the parentheses of the numerator and denominator of the representation of its probability as in (3.1) and can be derived by successively assuming zero partial associations among various variable pairs. Following Wermuth [10], a multiplicative model is constructed as follows: Given a multiplicative model, choose a variable pair (i_j,i_l), which is to have zero partial association, from a variable group in the numerator. Here (i_j,i_l) is a variable pair that is not contained in any one of the variable groups in the denominator. Denote by (i_j,i_l,i_K) the variable group in the numerator that includes (i_j,i_l), where i_K denotes the variables other than i_j and i_l. To get the desired model, we have only to replace $p(i_j,i_l,i_K)$ in the numerator by $p(i_j,i_K)p(i_l,i_K)$ and multiply the denominator by $p(i_K)$ and cancel the common factors. For instance, if

we assume the zero partial association of pair (i_1, i_4) in the above model, the application of the rule to (3.1) leads to the following model

$$\begin{aligned}(3.3)\qquad p(i_1,\cdots, i_5) &= p(i_1, i_5)p(i_4, i_5)p(i_2, i_4, i_5)p(i_3, i_4, i_5) \\ &\quad /\{p(i_5)p(i_4, i_5)p(i_4, i_5)\} \\ &= p(i_1, i_5)p(i_2, i_4, i_5)p(i_3, i_4, i_5)/\{p(i_5)p(i_4,i_5)\}\ .\end{aligned}$$

To search for an optimal combination of predictors of a variable i_1, we want to eliminate those variables which will show zero partial association with the variable i_1. For this purpose we define a particular sequence of models as follows:

$$\begin{aligned}&\text{MODEL}\,(0, 1):\ p(i_1,\cdots, i_k)=p(i_1,\cdots, i_k)\\ &\text{MODEL}\,(1, 1):\ p(i_1,\cdots, i_k)=p(i_1,\cdots, i_{k-1})p(i_2,\cdots, i_k)/p(i_2,\cdots, i_{k-1})\\ &\qquad(1, 2):\ p(i_1,\cdots, i_k)=p(i_1,\cdots, i_{k-2}, i_k)p(i_2,\cdots, i_k)\\ &\qquad\qquad\qquad /p(i_2,\cdots, i_{k-2}, i_k)\\ &\qquad\qquad\cdots\cdots\cdots\cdots\cdots\\ &\qquad(1, {}_{k-1}C_1):\ p(i_1,\cdots, i_k)=p(i_1, i_3,\cdots, i_k)p(i_2,\cdots, i_k)\\ &\qquad\qquad\qquad /p(i_3,\cdots, i_k)\\ &\text{MODEL}\,(2, 1):\ p(i_1,\cdots, i_k)=p(i_1,\cdots, i_{k-2})p(i_2,\cdots, i_k)/p(i_2,\cdots, i_{k-2})\\ (3.4)\qquad &(2, 2):\ p(i_1,\cdots, i_k)=p(i_1,\cdots, i_{k-3}, i_{k-1})p(i_2,\cdots, i_k)\\ &\qquad\qquad\qquad /p(i_2,\cdots, i_{k-3}, i_{k-1})\\ &\qquad\qquad\cdots\cdots\cdots\cdots\cdots\\ &\qquad(2, {}_{k-1}C_2):\ p(i_1,\cdots, i_k)=p(i_1, i_4,\cdots, i_k)p(i_2,\cdots, i_k)\\ &\qquad\qquad\qquad /p(i_4,\cdots, i_k)\\ &\qquad\qquad\qquad\vdots\\ &\text{MODEL}\,(k-2, 1):\ p(i_1,\cdots, i_k)=p(i_1, i_2)p(i_2,\cdots, i_k)/p(i_2)\\ &\qquad(k-2, 2):\ p(i_1,\cdots, i_k)=p(i_1, i_3)p(i_2,\cdots, i_k)/p(i_3)\\ &\qquad\qquad\cdots\cdots\cdots\cdots\cdots\\ &\qquad(k-2, {}_{k-1}C_{k-2}):\ p(i_1,\cdots, i_k)=p(i_1, i_k)p(i_2,\cdots, i_k)/p(i_k)\\ &\text{MODEL}\,(k-1, {}_{k-1}C_{k-1}):\ p(i_1,\cdots, i_k)=p(i_1)p(i_2,\cdots, i_k)\ .\end{aligned}$$

These model are generated by successively assuming zero partial associations and applying the above rule. MODEL (0, 1) means unconstrained model. MODEL (1, 1) represents the zero partial association between the variable i_1 and i_k in the sense that they are independent given the remaining $k-2$ variables. Similarly, MODEL (2, 1) represents the zero partial association between the variable i_1 and the set of variables $\{i_{k-1}, i_k\}$. Other models can be interpreted analogously. The variables appearing in the denominators of these equations define the candidates of the optimal combination of predictors of the variable i_1. The number 'l' of MODEL (l, m) denotes the number of zero partial associations

to be assumed and the number 'm' denotes that the model is the mth with respect to some proper ordering of the models belonging to the class of models with one and the same 'l'. Therefore, m does not exceed ${}_{k-1}C_l$. The total number of models belonging to the above sequence of models is given by

$$ {}_{k-1}C_0+{}_{k-1}C_1+\cdots+{}_{k-1}C_{k-1}=2^{k-1} . $$

For $k=2$ we get the unrestricted and the independence model discussed in Introduction.

Consider the set of variables defined by $I=\{i_2,\cdots,i_k\}$. Denote by E a subset of I. Taking into account that MODEL (0, 1) can be written as $p(i_1,\cdots,i_k)=p(i_1,\cdots,i_k)p(i_2,\cdots,i_k)/p(i_2,\cdots,i_k)$, or, using the above notations, $p(i_1, I)=p(i_1, I)p(I)/p(I)$, a model in the above sequence (3.4) can be represented in the form

$$ p(i_1, I)=p(i_1, E)p(I)/p(E) , \tag{3.5} $$

where we assume that $p(E)=1$ for $E=\phi$, an empty set. The AIC for the model (3.5) is given by

$$ \begin{aligned} \mathrm{AIC}=(-2)\sum_{i_1, I} n(i_1, I)\log\left[n(i_1, E)n(I)/\{n\cdot n(E)\}\right] \\ +2\{(C_{i_1}C_E-1)+(C_I-1)-(C_E-1)\} , \end{aligned} \tag{3.6} $$

where C_E and C_I denotes the number of categories of the corresponding sets of variables and we assume that $n(E)=n$ and $C_E=1$ for $E=\phi$. In calculating AIC's it is assumed that $0\log 0=0$. For the purpose of comparison of models within the above sequence, the common constant $(-2)\sum_I n(I)\log\{n(I)/n\}+2(C_I-1)$ can be ignored and the AIC is given by

$$ \mathrm{AIC}=(-2)\sum n(i_1, E)\log\{n(i_1, E)/n(E)\}+2\{(C_{i_1}C_E-1)-(C_E-1)\} . \tag{3.7} $$

This shows that we can compare these models without using the full-dimensional table. Further we note that from the point of view of AIC the comparison of models belonging to (2.2) reduces to that of

Table 3 Which sex

	S_1															
	A_1				A_2				A_3				A_4			
	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
W_1	50	57	77	55	38	58	61	51	38	39	59	58	42	40	95	86
W_2	28	27	40	26	17	28	40	35	16	20	39	37	21	23	45	49
Total	78	84	117	81	55	86	101	86	54	59	98	95	63	63	140	135

* See Tables 1.1, 1.2 and 1.3 about notations.

the models, MODEL $(k-2, m)$, $m=1, \cdots, {}_{k-1}C_{k-2}$, of (3.4) since the statistic (3.7) is identical to (2.4) when $E=\{i_l\}$, $l=2, \cdots, k$.

Table 3 is the four-way contingency table of the question 'which sex has more difficult life' and the three demographic factors. It will be interesting to see what combination of predictors the MAICE procedure adopt as the optimal one. The necessary eight models and their AIC's are obtained by putting $k=4$ in (3.4) and using (3.6). The results are given in Table 2. The MAICE is MODEL (2, 1) and shows that still a single factor sex defines the best combination to define the predictor. The result of Table 2 gives a finer description of the interdependence relation between the opinion and other demographic predictors than the result of the simplified analysis of the preceding section. Nevertheless, the result shows that we have only to pay our attention to sex in the case of the analysis of the interaction between the opinion and other demographic predictors.

The survey of Japanese national character has been conducted every five years since 1953. We used questionnaires which were common to all five surveys for the purpose of detection of changes in people's way of thinking. We applied the MAICE procedure proposed in this paper to the analysis of all questions of the 1973 survey. The results are quite assuring. In almost all the cases the MAICE lead to the same conclusion as that obtained by a careful analysis of the data formerly reported in [6].

The analysis of a multidimensional contingency table has been a difficult and very much time-consuming task. This was mainly due to the inappropriate modeling and the lack of an objective criterion for the evaluation of the badness of a fitted model. By applying the procedure of this paper we can easily find what combination of predictors is the most important as a factor and list up the predictors in order of the dependence of the variable on the predictors. The use of the statistic (3.7) also facilitate the search for the optimal combination of predictors for a high dimensional table. This last aspect will be discussed in more detail in a future paper.

has more difficult life?

S_2																Total
A_1				A_2				A_3				A_4				
R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4	
47	57	63	48	40	50	57	53	32	45	47	45	35	42	66	63	1694
32	56	63	52	29	57	67	46	37	44	61	49	37	49	90	101	1361
79	113	126	100	69	107	124	99	69	89	108	94	72	91	156	164	3055

4. Discussion of statistical characteristics of the procedure

Suppose that a four-dimensional probability distribution is defined by

$$p(i_1, i_2, i_3, i_4)=p(i_1, i_2)p(i_2, i_3, i_4)/p(i_2) .$$

We assume the values of these probabilities shown in Table 4. The question is whether we can detect the true structure by the MAICE procedure. To answer this question we generated 100 sets of data each composed of 3000 random samples from the above probability distribution. The frequencies of the models chosen as the MAICE's are shown

Table 4

i_1, i_2	Probability	i_1, i_2	Probability
1, 1	0.2959	2, 1	0.1607
1, 2	0.2586	2, 2	0.2848

i_2, i_3, i_4	Probability	i_2, i_3, i_4	Probability
1, 1, 1	0.0255	2, 1, 1	0.0259
1, 1, 2	0.0275	2, 1, 2	0.0370
1, 1, 3	0.0383	2, 1, 3	0.0412
1, 1, 4	0.0265	2, 1, 4	0.0327
1, 2, 1	0.0180	2, 2, 1	0.0226
1, 2, 2	0.0281	2, 2, 2	0.0350
1, 2, 3	0.0331	2, 2, 3	0.0406
1, 2, 4	0.0282	2, 2, 4	0.0324
1, 3, 1	0.0177	2, 3, 1	0.0226
1, 3, 2	0.0193	2, 3, 2	0.0291
1, 3, 3	0.0321	2, 3, 3	0.0353
1, 3, 4	0.0311	2, 3, 4	0.0308
1, 4, 1	0.0206	2, 4, 1	0.0236
1, 4, 2	0.0206	2, 4, 2	0.0298
1, 4, 3	0.0458	2, 4, 3	0.0511
1, 4, 4	0.0442	2, 4, 4	0.0537

Table 5

Estimated Distribution	Frequency	Number of Free Parameters	Frequency Accepted by χ^2-test	Degrees of Freedom
$p(i_1, i_2, i_3)p(i_2, i_3, i_4)/p(i_2, i_3)$	6	39	95	24
$p(i_1, i_2, i_4)p(i_2, i_3, i_4)/p(i_2, i_4)$	9	39	97	24
$p(i_1, i_2)p(i_2, i_3, i_4)/p(i_2)$	85	33	95	30
Other distributions	0	—	0	—
Total	100	—	—	—

in Table 5. The result tells that the MAICE procedure produced correct answer 85 times out of 100. Needless to say the performance of the procedure depends on the sample size and the structure of the true distribution. The present result is a typical example as is expected from the definition of AIC statistic.

Consider that the chi-square goodness of fit tests are applied to our example. We regard the situation as the fitting of each of the models described in Section 3 to the observations $n(i_1, i_2, i_3, i_4)$. For example, for the above model we get the chi-square test statistic

$$\chi^2=\sum_{i_1,\cdots,i_4}\{n(i_1, i_2, i_3, i_4)-n(i_1, i_2)n(i_2, i_3, i_4)/n(i_2)\}^2 \\ /\{n(i_1, i_2)n(i_2, i_3, i_4)/n(i_2)\}\ .$$

The figures in the right half of Table 5 give the frequency for each model accepted at the level of 5%. The results show that three models including the true one were accepted about 95 times out of 100. This means that the test procedure can not discriminate more complicated models from the true structure.

The figures in the right half of Table 2 give χ^2, $1-F(\chi)^2$ and the degrees of freedom for each of the seven models of the data given by the four-way contingency table shown in Table 3. Here F denotes a cumulative distribution function of a chi-square variable. If the test is applied only to those models within MODEL $(2, m)$, the χ^2 for the MODEL $(2, 1)$ is insignificant, with respect to the 5% level of significance. This result shows that the model is acceptable, or at least not rejected, and coincides with the conclusion by MAICE for this case. However, if the test is applied to models defined by $(2, 2)$, every model is rejected at the level of 5%, as is shown in the three lines from the bottom of Table 2. The MAICE is Model No. 1° for this case too, but by the test procedure MODEL $(0, 1)$ is the only choice.

The relation between the MAICE and classical test procedures can be understood by considering the fact that the log likelihood ratio test statistic takes the form $\chi^2=\mathrm{AIC}\,(k)-\mathrm{AIC}\,(K)+2(K-k)$, where $\mathrm{AIC}\,(k)$ denotes the AIC of a model with k free parameters. K is usually the highest possible value of k and χ^2 is tested as a chi-square with the degrees of freedom $\mathrm{d.f.}=K-k$. Taking into account that the expectation of χ^2 is equal to its degrees of freedom, we can understand that the MAICE procedure applied to each pair of models in the above example means the comparison of the value of χ^2 with twice its expectation. The values of $1-F(2\,\mathrm{d.f.})$ for various values of d.f. are given in Table 6. The table clearly shows that by AIC the "level of significance" is adjusted in such a way that the corresponding probability of rejection of the simpler model decreases as the degrees of freedom

Table 6

d.f.	$1-F$(2d.f.)	d.f.	$1-F$(2d.f.)
1	0.1572989537	10	0.0292526881
2	0.1353352832	15	0.0119215009
3	0.1116101347	20	0.0049954123
4	0.0915781944	25	0.0021311519
5	0.0752352001	30	0.0009206824
6	0.0619688044	40	0.0001763029
7	0.0511816101	50	0.0000345493
8	0.0423801120	60	0.0000068763
9	0.0351737134	70	0.0000013839

* Calculated by expansion formula for the χ^2-distribution function

increase. The MAICE procedure, therefore, has a tendency to adopt simpler models compared with the chi-square test procedure as the degrees of freedom increase. This characteristic of the MAICE seems to be in better agreement with our intuitive choice when a complex model is fitted than the one by the chi-square test. Now if a modification of a test procedure considered so that the significance level is adjusted in accordance with the degrees of freedom, one has to provide a rule for the adjustment. Even if this adjustment is made possible, it is still impossible to compare a model with every possible choice of the alternative. For example, it is impossible to compare MODEL (1, 1) with MODEL (1, 2) in Table 2 by the classical chi-square test. The salient feature of AIC is that it is an estimate of a clearly defined universal measure of fit, the entropy defined in Section 1. This fact justifies the comparison of AIC's among every possible model which cannot necessarily be compared by the classical goodness of fit test.

5. Concluding remarks

Generally there are two different types of analysis of survey results. The one is the case where the purpose of the analysis is to evaluate the dependence between a specific variable to be predicted and a specific predictor, such as the answer to the question " Which political party the youth has been supporting? " The other is the case where the object is to seek an explanation of phenomenon, exemplified by the question " What has caused the changes in political party support? " We are sure that the procedure proposed in this paper will be of great help to solve the latter problem. By our procedure, as was shown in preceding sections, the comparison of various models is very simple and under certain circumstances the search procedure for the optimal combination of predictors can be done without the use of the full-

dimensional contingency table.

The definition of AIC will draw researcher's attention to the relation between the number of free parameters within a model and the sample size of the survey data. This aspect of statistical analysis was not clearly recognized in the application of classical tests. We are tempted to think that classical tests, such as the chi-square test of goodness of fit and that of independence, are merely approximate realizations of our procedure. However, our procedure needs further refinement of the basic model so as to take care of the situation where many cells are lacking observations. This will be the subject of further study.

A Fortran program for the entire procedure is available from the authors.

Acknowledgement

We wish to thank the referees for their helpful comments. Thanks are due to Mr. K. Katsura of the Institute of Statistical Mathematics for programming and testing the algorithms. We are grateful to Mr. G. Kitagawa and Mr. M. Ishiguro for their helpful suggestions.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki, Eds., Akademiai Kiado, Budapest, 267-281.

[2] Akaike, H. (1976). On entropy maximization principle, *Applications of Statistics*, P. R. Krishnaiah, ed., North-Holland, Amsterdam, 27-41.

[3] Bishop, Y. M. M. (1969). Full contingency tables, logits and split contingency tables, *Biometrics*, **25**, 383-400.

[4] Darroch, J. N. (1962). Interactions in multifactor contingency tables, *J. R. Statist. Soc.*, B**24**, 251-263.

[5] Goodman, L. A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications, *J. Amer. Statist. Ass.*, **65**, 226-256.

[6] Research Committee on the Study of Japanese National Character (1976). *Nipponjin no Kokuminsei, sono san* (A study of the Japanese National Character, Part III), Shiseido, Tokyo. (In Japanese)

[7] Sakamoto, Y. (1974). A study of the Japanese National Character—Part V, *Ann. Inst. Statist. Math., Supplement* 8, 1-58.

[8] Sakamoto, Y. (1977). A model for the optimal pooling of categories of the predictor in a contingency table, *Research Memorandum*, No. 119, The Institute of Statistical Mathematics, Tokyo.

[9] Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance analysis, *Biometrics*, **32**, 95-108.

[10] Wermuth, N. (1976). Model search among multiplicative models, *Biometrics*, **32**, 253-263.